EXHIBIT L

Libgen - Anna Archive

Oncall Vladan Petrovic (gaid) · Modality Text Image · Category GenAI

Add to Collection | Hive Deletion | Activity Log

Overview | Compliance | Lineage | Explore | Model Snapshots

Privacy Review Status

Dataset Review Status: Completed | Review Required By: No update time info | Priority: None | Final Decision: Passed

Privacy & Crawling

This card should be filled by a **fact gathering specialist** only!

Old Dataset Facts

Outline

- Basic Information
- License
- Construction
- Dataset Access
- Youth
- Content
- Youtube Analysis
- Other

Basic Information

Full/Other Names	Insert full or other names
Dataset Homepage	Insert dataset homepage
Introducing Paper	Insert introducing paper
Publisher	Insert publisher
Summary	Insert summary

License Information

Final Decision

Final Decision: Passed

Approved Usages: Training

Complete Final Decision | Send Back

Requirements

- Face Anonymization: I will blur all faces in the dataset unless verified that no individuals from Illinois or Texas are included.
- Access Control: I will use appropriate Access Control Lists (ACLs) to restrict data access to those with a legitimate business need.
- Crawling Compliance: If using automation to download the dataset, I will filter URLs based on the blocklist and adhere to directives in Robots.txt and "No AI" tags.
- Prohibition of CSAM: I will ensure no child sexual abuse material (CSAM) is used for training models.

License Types	Select license types
License Notes	Insert license notes

Construction Information

Construction Methods	Select construction methods
Data Origin Source	Insert data origin source
Domain Blocklist	Insert domain blocklist here
TV Shows, Movies, Stock Photos	Insert included media here
Construction Notes	Insert construction notes

If Crawled

Meta or 3P Crawled	Insert meta or 3P crawl here
Login Required?	Insert login required here

Data Retention Limit: I will not retain data for longer than 1095 days.

Handling of 1PD in 3PD Dataset: If I detect first-party data (1PD) in the 3PD dataset, I will cease its use and consult the Dataset Review Team.

Does Robots.txt Restrict Access?	Insert robot restrict access here
Content or IDs/URLs?	Insert crawl content type here
If Crawled or Combined from Existing:	
Full Curation / Annotation?	Insert full curation or annotation here
Content Fully from Trusted Sources?	Insert is trusted source here
If Annotated / Derived / Combined from Existing:	
Source Dataset(s)	Insert source datasets
If Production:	
Does this 3P dataset contain 1PD?	Insert contains 1PD here
If Synthetic:	
Does dataset contain synthetic data?	Insert is contains synthetic data here

Dataset Access Information

Access Method	Insert access

	method here
Access Notes	Insert access notes

Youth

Contains Youth data (u18)?	Choose Yes/No
Is Youth (u18) data the focus or incidental?	Insert data presence here

Text Content Information

Contains Text?	Choose Yes/No
Is there PII present in the dataset?	Choose Yes/No
If PII is present, is it incidental or the focus of the dataset?	Insert data presence here
Is News content present in the dataset?	Choose Yes/No
If News content is present, is it incidental or the focus of the dataset?	Insert data presence here
Text Content Notes	Insert text content notes

Visual Content Information

Contains Images or Video?	Choose Yes/No

Contains People in Images or Video?	Insert contains people here
Do we know provenance of images or video?	Insert provenance image or video here
IL/TX for People in Images or Video?	Insert image and video IL/TX here
Image or Video Annotations?	Insert images or videos annotation here
Visual Annotation Type(s)	Select visual annotation types
Image or Video Annotations Involve Humans?	Insert annotation involve human here
Visual Content Notes	Insert visual content notes

Audio Content Information

Contains Audio?	Choose Yes/No

Contains Music?	Insert contains music here
Contains Music Lyrics? (text)	Insert contains music lyrics here
Audio Annotations?	Insert contains audio annotations here
Audio Annotations Type	Insert audio annotations type here
Human Voices Present Within Audio?	Insert human voices presented here
IL/TX for Speaking People?	Insert is there IL/TX for Speaking People here
Audio Content Notes	Insert audio content notes

YouTube Analysis

Publisher Y/N	Insert youtube

	publisher
# Citations	Insert number of citations
Citations Link	Insert citations link
Top 15 Citation(s)	Insert top 15 citations
Citations	Insert is citations
Info	Insert google info
Is Google Crawled	Insert is google crawled
Industry Standard Final Determination	Insert industry standard final determination here
Industry Standard?	Insert industry standard
Chinese Origin?	Choose Yes/No
If Chinese Origin:	
Government Suppliers	Insert government

	suppliers here
Surveillance Data	Insert surveillance data here
Sensitive Topics and Subject Matter Domains	Insert sensitive topics and subject matter domains here
Is Present on EU/US Notorious Markets List or Piracy Lists	Choose Yes/No
File names include 'Pirated' or 'Stolen'	Choose Yes/No
Reputation for hosting or providing Pirated material	Choose Yes/No
Restricted Data Present	Insert restricted data presence here
CBRNE Content Present	Insert CBRNE content presence here

Other

Other notes	Insert other notes

Activity & Comments

Type a comment

Morphing Framework Bot added source: https://annas-archive.org/datasets/libgen_rs
February 16, 10:34 AM · Like

Morphing Framework Bot updated modalities to [text, image]
January 8, 12:47 PM · Like

Morphing Framework Bot created this AIDC Dataset.
December 12, 2024 · Hide updates

added the tag genai.

created this AIDC Dataset.

TDM Processor Completed the stage Requirements
September 16, 2024 · Like

Christian Montez updated the "Final Decision" to "PASSED"
July 30, 2024 · Like

Alisa Hall made several changes
July 29, 2024 · Hide updates

updated the "Product Counsel Approval Status" to **Redacted - Privilege**

updated the "Product Counsel Notes" to **Redacted - Privilege**

Christian Montez made several changes

July 29, 2024 · Hide updates

updated the "Privacy Notes" to "High risk dataset - new use cases / usage in new models requires additional review and director level approval"

updated the "Privacy Risk Level" to "High"

updated the "Privacy Approval Status" to "Approved"

updated the "Other" to "IP Policy Analysis:

As IP Policy has previously noted with respect to GenAI training on content from sites such as LibGen that contain pirated content, policymakers will take a negative view of such sites and their use (irrespective of any legal concerns), and such concerns may contribute to spurring future regulatory efforts and/or complicate our efforts to build goodwill with governments on AI regulation, especially in the US & Europe. In addition to these previously-flagged policy risks, several recent developments heighten the policy risks:

In June, OpenAI made public a statement (https://www.copyright.gov/policy/artificial-intelligence/ex-parte-communications/letters/OpenAI-June-4-2024.pdf) that it avoids crawling “sites that are known to engage in IP infringement, such as the ‘notorious markets’ identified annually by the Office of the U.S. Trade Representative.”
Mark has recently been engaging with high-level U.S. government officials on LLaMA and our AI strategy, including Sec. Raimondo and NSA Jake Sullivan, and because of OpenAI’s statement, this is now a question that could arise in such engagements (indeed, Commerce in particular, takes strong stands on IP piracy matters).
In a hearing last December, Rep. Deborah Ross (D-NC) questioned Matt Schruers , the head of our trade association CCIA about his members’ use of pirate web sites, including Sci-Hub, for training AI.
The growing popularity of transparency obligations, from the EU AI Act to bills introduced in Congress and elsewhere, suggests that what we train on will be public sooner rather than later.
The fact that we train on pirated sites such as Anna’s Archive is likely to be viewed negatively by policymakers, rightsholders, and other stakeholders.

POLICY VIEW AND RECOMMENDATION:
On balance, Policy doesn’t view these developments as categorically changing the risks flagged earlier, and thus at this time, we are not seeking to reopen the prior decision w.r.t the content from Anna’s archive.
To mitigate these emerging policy risks, however, we recommend that we commit to not training on other “notorious sites (https://ustr.gov/sites/default/files/2023_Review_of_Notorious_Markets_for_Counterfeiting_and_Piracy_Notorious_Markets_List_final.pdf)” besides Anna’s archive.
"

Luc Dahlin made several changes

July 24, 2024 · Hide updates

updated the "Ip Legal **Redacted - Privilege**

updated the "Ip Legal [redacted]
updated the "Ip Legal Approval Status" to [redacted]
updated the "Ip Legal Notes" to [redacted]

[redacted]

Christian Montez made several changes
July 12, 2024 · Hide updates

updated the "Final Decision" to "BLOCKED_DUE_TO_ESCALATION"
updated the "Review Status" to "COMPLETED"

Fatima Jafri made several changes
July 12, 2024 · Hide updates

updated the "Product Counsel [redacted]
updated the "Product Counsel Notes" to [redacted]
updated the "Product Counsel Approval Status" to [redacted]

Tyler Robbins updated the "Ip Legal Approval Status" to [redacted]
July 5, 2024 · Like

Luc Dahlin updated the "Ip Legal Notes" to "Note for product counsel: [redacted]

[redacted]

July 5, 2024 · Like

Tyler Robbins made several changes

June 27, 2024 · Hide updates

- updated the "Ip Legal Notes" to Redacted - Privilege

Redacted - Privilege

- updated the "Ip Legal Redacted - Privilege
- updated the "Ip Legal Notes" to "[TYLER] Redacted - Privilege

Redacted - Privilege

- updated the "Ip Legal Approval Status" to Redacted - Privilege

David McAneny made several changes

June 11, 2024 · Hide updates

- updated the "Review Status" to "LEGAL_REVIEW"
- updated the "Login Required For Download Text" to "Full datasets require download via torrent
  Can also search one off books, articles, etc on the website that can be dowloaded"
- updated the "Login Required For Download Flag" to "false"
- updated the "Type Of Pii" to "Full names, email"
- updated the "Pii Present Flag" to "true"
- updated the "Pii Present Text" to "Each dataset is hundreds of GB or TB of data, unable to download due to extreme size of datasets, in addition, full downloadable datasets are accessible via torrent, which is blocked
  Looking at sample metadata on the website indicates that it is possible for full names, and emails to be contained at least"
- updated the "Scraping Prohibited In Tos Text" to "Claims all code and data is open source
  Unsure of copyright requirements/obligations for each book, paper, etc. hosted on the website"
- updated the "Scraping Prohibited In Tos Flag" to "false"

updated the "Robots Txt Disallow Flag" to "true"

updated the "Robots Txt Disallow Text" to "User-agent: *
Crawl-delay: 10
Disallow: /db
Disallow: /slow_download
Disallow: /fast_download
Disallow: /torrents
Disallow: /search
Disallow: /scidb"

updated the "Attributes Required Text" to "If you are interested in mirroring this dataset for archival or LLM training purposes, please contact us.

Claims all code and data is open source.

Don't need to credit anna's archive, however unsure of copyright or attribution requirements/obligations for each book, paper, etc. hosted on the website"

updated the "Contain 1pd Text" to "Each downloadable dataset is hundreds of GB or TB of data, unable to download due to extreme size of datasets. Did not observe 1PD in the observable data"

updated the "Contain 1pd Flag" to "false"

updated the "Human Characterization Text" to "Each downloadable dataset is hundreds of GB or TB of data, unable to download due to extreme size of datasets, as a result unable to determine full scope.

Searching on the website shows that there are human faces associated with full names"

updated the "Human Characterization Flag" to "true"

updated the "Attributes Required Flag" to "true"

updated the "Attributes Required Text" to "If you are interested in mirroring this dataset for archival or LLM training purposes, please contact us.

Claims all code and data is open source"

updated the "Scraping Notes" to "Sample metadata indicates that links from various news sources like NYTimes is contained"

updated the "Review Status" to "IN_PROGRESS"

Praveen Krishnan added the tag genai.
May 29, 2024 · Like

Praveen Krishnan changed the oncall to gaid

May 29, 2024 · Like

Praveen Krishnan changed the owner to Praveen Krishnan.
May 29, 2024 · Like

Praveen Krishnan set ACL to Gen_AI
May 29, 2024 · Like

Praveen Krishnan changed the title to "Libgen - Anna Archive ".
May 29, 2024 · Like

Praveen Krishnan created this Dataset Catalog Dataset Custom Fields.
May 29, 2024 · Hide updates

- updated the "Used Modality" to "Image, Text"
- updated the "Approved Usages" to "Training"
- updated the "Review Status" to "SUBMITTED"
- updated the "Llama4 Review Task" to "375210415523490"
- updated the "Model Family" to "LLAMA4"